Human Review Has to Be Tested

“A human checked it” describes a step in a workflow. To rely on that step, an organization needs evidence that the reviewer can find consequential errors and act on them.

Human Review Has to Be Tested

An AI-assisted evidence brief arrives with a reassuring sentence: “Reviewed by an analyst.” The document is fluent, its tables look consistent and its citations resolve. A manager approves it for use in a funding discussion.

What did the analyst actually check?

They may have compared every important claim with its source. They may have checked a sample. They may have read for plausibility under a deadline. The same approval label can conceal very different controls.

A workable review process specifies the evidence the reviewer sees, the errors they are expected to catch, the time available and what happens when something is wrong. It should also be tested under conditions that resemble the work.

© 2026 Prova LLC · All rights reserved · PP-000241-10

01 Start with an error proofreading will miss

The people, places, organizations and figures in this document's examples are fictional.

Consider an example. A brief says that a hospital-transition program reduced readmissions by 18 percent in a randomized trial. The number appears in the underlying report, but it is a relative reduction from a subgroup of participants who completed the service. The primary assignment-effect estimate is smaller and imprecise.

A reviewer who checks only that the number and citation exist may approve an overextended claim. Detection requires finding the relevant analysis, distinguishing assignment from completion and checking the population, comparison and uncertainty.

The NIST Generative AI Profile identifies risks involving both false but plausible outputs and human–AI configurations, including overreliance. That supports examining the whole arrangement: task, source material, interface, model and reviewer.

For this task, the review artifact should put the claim beside the source table or passage, its page number, the estimate type and any transformation. A paragraph-level citation may not provide enough information to perform the check efficiently.

© 2026 Prova LLC · All rights reserved · PP-000241-10

02 Choose the right comparison for the workflow

Human involvement does not guarantee a performance improvement. But evidence about that point needs careful interpretation.

Vaccaro, Almaatouq and Malone's 2024 meta-analysis covered 106 experiments and 370 effect sizes. On average, human–AI combinations performed better than humans alone, but worse than the better of humans or AI alone. These are different comparisons. The review also found substantial variation across tasks and studies; it does not establish how a particular evidence-review workflow will perform.

The operational question is whether the proposed arrangement improves the outcomes that matter compared with the existing process. Those may include the rate of consequential errors, the time to a checked conclusion, consistency across reviewers and whether important groups or findings are overlooked. Faster drafting alone cannot answer that question.

A 2025 survey by Lee and colleagues collected 936 examples from 319 knowledge workers. Higher confidence in generative AI was associated with less reported critical thinking, while higher task-specific self-confidence was associated with more. This was a self-report survey, not an experiment showing that AI caused reasoning skills to deteriorate. It identifies a reason to investigate reliance, not a verdict on every use of AI.

© 2026 Prova LLC · All rights reserved · PP-000241-10

03 Specify what makes review possible

A review control needs six practical elements.

An error model. Name the failures that matter: wrong population, wrong outcome, a relative effect presented as absolute, a subgroup result promoted to the headline, a transcription error or an unsupported causal verb. “Check accuracy” leaves too much unspecified.

Evidence at the point of review. Supply the source material and transformations needed to reproduce consequential claims. An AI-generated explanation of its answer cannot substitute for the underlying evidence.

Appropriate expertise. Verifying a transcription, appraising an identification strategy and interpreting a service pathway require different knowledge. Match the reviewer to the judgment rather than assigning everything to a generic approval role.

Time and attention. If output grows tenfold while review capacity stays fixed, the control has changed. Account for checking time and set coverage according to the task and consequence. Sampling needs a defined rule and an escalation threshold.

Authority to act. State who can reject a result, require manual rework, widen the check or pause the workflow. Comments that cannot affect release provide limited protection.

A record of what happened. Record the error, its severity, the relevant configuration and its disposition. Use repeated failures to change the workflow and retest it.

This structure is Prova’s practical application of established assurance principles. The GAO AI Accountability Framework organizes accountability around governance, data, performance and monitoring. The Magenta Book, sections 5.6 and 5.9, calls for verifiable AI-assisted evaluation work and quality assurance planned across design, execution, analysis and reporting.

© 2026 Prova LLC · All rights reserved · PP-000241-10

04 Test whether the review catches the problem

A written specification describes an intended control. The next step is to observe it working.

One useful model appears in HM Treasury's guidance on evaluating AI interventions. Its **hypothetical** briefing-note case combines time measurement with a predefined quality assessment, including an independent assessor who is not told whether AI was used. It also proposes examining whether false model outputs survive into final briefs. The example illustrates an evaluation design; it is not evidence of an observed productivity gain.

For an evidence-extraction workflow, Prova's suggested pilot is to prepare a representative set of records with adjudicated reference answers. Include clean cases, ambiguous cases and known consequential errors. In a controlled test set, some errors can be deliberately introduced to test detection; keep these out of operational evidence products. Report the mixture of deliberately introduced and naturally occurring errors; detection in an error-enriched test set does not estimate the live workflow's error prevalence.

Have reviewers use the intended interface, source access and time allowance. Measure what they detect, what they miss, which correct findings they incorrectly reject and whether important judgments vary between reviewers. Record checking and rework time as well as drafting time.

The reference answers also need scrutiny. A second competent reviewer should examine consequential cases, and genuine disagreements should remain visible. A test cannot validate a workflow against an unreliable answer key.

Where feasible, compare the assisted workflow with the existing workflow on comparable cases. Establish what improvement would justify use and which missed errors would require redesign. The acceptable threshold depends on consequence; this article does not prescribe a universal error rate.

© 2026 Prova LLC · All rights reserved · PP-000241-10

05 Keep checking as the system changes

Review performance is a property of a particular arrangement. A model update, new source format, larger batch, different reviewer or shorter deadline may alter it.

Keep enough information to identify what was tested: model or tool version where available, task instructions, source preparation, interface, reviewers and decision rules. Recheck consequential changes and monitor errors that appear during real use. A pilot result should not become an indefinite permission slip.

There is a practical limit. If a claim cannot be checked at the proposed speed or with the available expertise, narrow the task, improve the source structure or retain manual analysis for that part of the work. Human responsibility becomes useful when the process gives the responsible person a realistic chance to exercise it.

Before increasing production, select one error that would change a real decision and watch the review process try to catch it. **The approval box matters only insofar as the work behind it is dependable.**

RELATED: The Trial Tested the Offer, for the assignment-versus-completion error used in the example; A Systematic Review Is a Study, for the source and decision records a reviewer needs to inspect.

© 2026 Prova LLC · All rights reserved · PP-000241-10

06 Further reading

1 HM Treasury, *The Magenta Book* (2026), sections 5.6 and 5.9, pp. 128–135 — AI-assisted evaluation and ongoing quality assurance.

2 HM Treasury, *Guidance on the Impact Evaluation of AI Interventions*, section 2.3 and Annex A, practical example 2 — baseline design and a hypothetical evaluation of briefing-note production. The linked HTML version was updated in May 2026.

3 NIST, *Generative Artificial Intelligence Profile* (2024) — risks including confabulation and human–AI configuration.

4 U.S. GAO, *Artificial Intelligence: An Accountability Framework* (2021) — governance, data, performance and monitoring.

5 Vaccaro, Almaatouq and Malone, "When combinations of humans and AI are useful" (2024) — distinct comparisons with humans alone and with the better solo performer; substantial variation across experiments.

6 Lee et al., "The Impact of Generative AI on Critical Thinking" (2025) — reported critical-thinking practices and associations with confidence.

QUESTION IT

Found an error, or want to question this document? Contact Prova at hello@byprova.com.

© 2026 Prova LLC · All rights reserved

Cite as: Prova (2026). Human Review Has to Be Tested. byprova.com

Examples are fictional unless sourced. Not legal or financial advice.